A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE

Available online at www.sciencedirect.com
ScienceDirect
Genomics 91 (2008) 281–288
GENOMICS
www.elsevier.com/locate/ygeno

A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE

Geoffrey J. Faulkner [a,*], Alistair R.R. Forrest [b,c], Alistair M. Chalk [c], Kate Schroder [a], Yoshihide Hayashizaki [b], Piero Carninci [b,d], David A. Hume [e], Sean M. Grimmond [a]

[a] The Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia
[b] Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan
[c] The Eskitis Institute for Cell and Molecular Therapies, Griffith University, Brisbane, QLD 4111, Australia
[d] Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, Wako, Saitama 351-0198, Japan
[e] Roslin Institute, University of Edinburgh, Roslin EH25 9PS, Scotland, UK

Received 15 August 2007; accepted 7 November 2007

Abstract

Cap analysis gene expression (CAGE) is a high-throughput, tag-based method designed to survey the 5′ end of capped full-length cDNAs. CAGE has previously been used to define global transcription start site usage and monitor gene activity in mammals. A drawback of the CAGE approach thus far has been the removal of as many as 40% of CAGE sequence tags due to their mapping to multiple genomic locations. Here, we address the origins of multimap tags and present a novel strategy to assign CAGE tags to their most likely source promoter region. When this approach was applied to the FANTOM3 CAGE libraries, the percentage of protein-coding mouse transcriptional frameworks detected by CAGE improved from 42.9 to 57.8% (an increase of 5516 frameworks) with no reduction in CAGE to microarray correlation. These results suggest that the multimap tags produced by high-throughput, short sequence tag-based approaches can be rescued to augment greatly the transcriptome coverage provided by single-map tags alone.

Keywords: Transcriptome; Promoter; CAGE; Microarray

Global analyses of the mammalian transcriptome have to date focused upon the synthesis, sequencing, and analysis of large-scale cDNA libraries [1–5]. The parallel development of high-throughout technologies designed to survey transcriptome structure and function efficiently has proven to be essential for these efforts. One such technology, cap analysis gene expression (CAGE), was introduced by Carninci et al. and implemented broadly by the third stage of the Functional Annotation of Mouse project (FANTOM3) [1].

CAGE interrogates the initial 20 or 21 nucleotides of capped, full-length cDNAs via tags cleaved from their 5′ end by the restriction enzyme MmeI [6–9]. When mapped to genomic sequences, these tags can be used to identify transcription start sites (TSSs) and—if sequencing is undertaken to a sufficient depth—quantify transcript expression. Genome-scale application of CAGE in human and mouse tissues and cell types has contributed greatly to the elucidation of mammalian promoter architecture and revealed that the majority of mammalian promoters do not initiate from a single TSS, but rather initiate from clusters of sites in 50-to 100-bp windows [10–14].

A major CAGE-related problem noted by FANTOM3 was that only 61.8% of the ~11.6 million FANTOM3 mouse CAGE tags could be mapped unequivocally to a single genomic location. The remaining tag complement either did not map or mapped to multiple genomic locations. For the purposes of FANTOM3, only single-map tags were included for further analyses due to the suggested ambiguity and noise inherent in multimap tags. Subsequently, the abundance of multimap CAGE tags has not been thoroughly addressed apart from suggestions that they are caused by CpG-rich promoters, genome-wide repeat elements, or a highly conserved TSS-proximal motif [1,9].

* Corresponding author. Fax: +61 7 3346 2101.
E-mail address: g.faulkner@imb.uq.edu.au (G.J. Faulkner).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.11.003

An alternative explanation for the multimap CAGE phenomenon is that short sequence tags extracted from the transcriptome are inherently far more redundant than random expectation would suggest, possibly as a consequence of gene duplication events. This concept could explain a troubling observation made by FANTOM3 in which, despite a broad range of mouse CAGE libraries (145 libraries sampling 23 cell lines and tissues at various activation states), one-third of known protein-coding mouse transcriptional frameworks (TKs) could not be assigned a TSS based on single-map CAGE tags alone.

Here we present evidence that (a) a significant proportion of the known transcriptome can be detected only by multimap CAGE tags; (b) multimap tags are a consequence of inherent redundancy in short sequence tags extracted from transcribed regions, rather than CpG islands, genome-wide repeat elements, or a conserved TSS-proximal motif; and (c) multimap tag rates are inversely proportional to tag length. Furthermore, we demonstrate that multimap tags can be reincorporated into the overall CAGE tag set through a novel strategy designed to select the most likely promoter region of origin for a given multimap tag. We then validate this method via a cross-platform comparison of CAGE to an Affymetrix array for both the original and the rescued tag sets. Last, we show that promoters predominantly detected by rescued multimap tags are far more likely to conform to a narrow, TATA-box-associated architecture than promoters detected by single-map tags alone.

Results and discussion

Characteristics of multimap CAGE tags

A revised genomic mapping of the 11,567,973 FANTOM3 mouse CAGE tags via Vmatch [15] revealed that 53.98% mapped uniquely to a single genomic location, 40.97% mapped to multiple locations, and 5.05% could not be mapped. The equivalent figures from the original FANTOM3 mappings were 61.8, 14.4, and 23.8%, respectively.

The differences in mapping rates were partly due to the use of Vmatch, a matching algorithm tailored toward the alignment of very short sequences to the genome, rather than BLAST [16], a heuristic algorithm not specifically designed for very short sequence matching. Our allowance for a single internal mismatch in tag-genome alignments—as well as 5′-and 3′-end mismatches—was likely responsible for the remaining differences in mapping rates compared with FANTOM3, for which mismatches were permitted only at the ends of tags.

An evaluation of the mapping distribution for all tags (Fig. 1) indicated that 90% of multimap tags mapped to 10 or fewer genomic locations. For the multimap portion of the distribution the median was 3 locations and the mode was 2 locations. Only a small fraction (2.4%) of tags mapped to more than 100 locations.

The origins of multimap CAGE tags

First and foremost, the contribution of CpG islands to the multimap tag phenomenon was examined by determining which tags originated from within the CpG islands defined by the UCSC Genome Browser [17]. If CpG islands—with their intrinsically reduced sequence complexity and common proximity to TSSs [11]—were disproportionately contributing to the production of multimap tags, the percentage of CpG-associated tags would increase with mapping location frequency. However, the percentage of CpG-associated tags actually decreased as mapping location frequency increased (Fig. 2), indicating that CpG islands do not disproportionately contribute to the production of multimap tags.

Next, a broader analysis of tag association with other repeat elements was performed using the RepeatMasker [18] coordinates available from the UCSC genome browser. As illustrated in Fig. 3, the percentage of multimap tags that mapped to 2–10

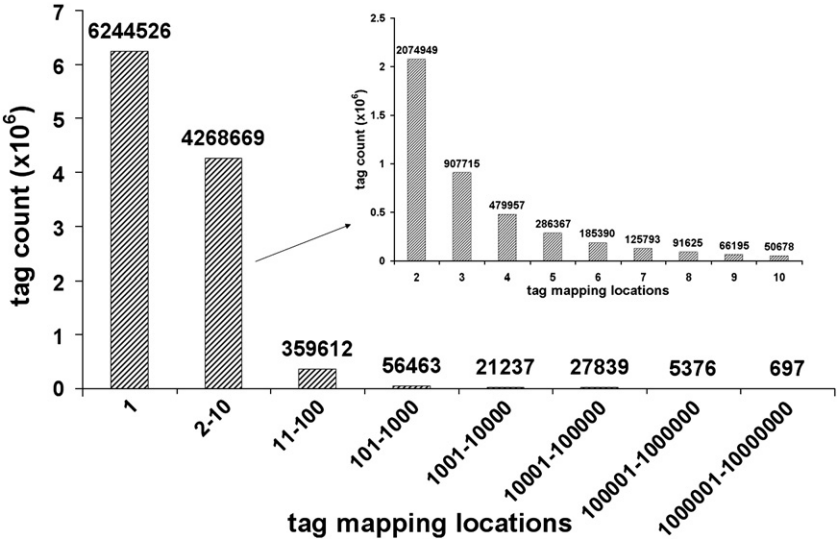

Fig. 1. CAGE tag-mapping distribution. Each column corresponds to the number of tags mapping to the specified number of genomic locations. A subdistribution for tags mapping to 2–10 locations is also shown (inset) to illustrate that the majority of multimap tags originated from 2 or 3 genomic positions.

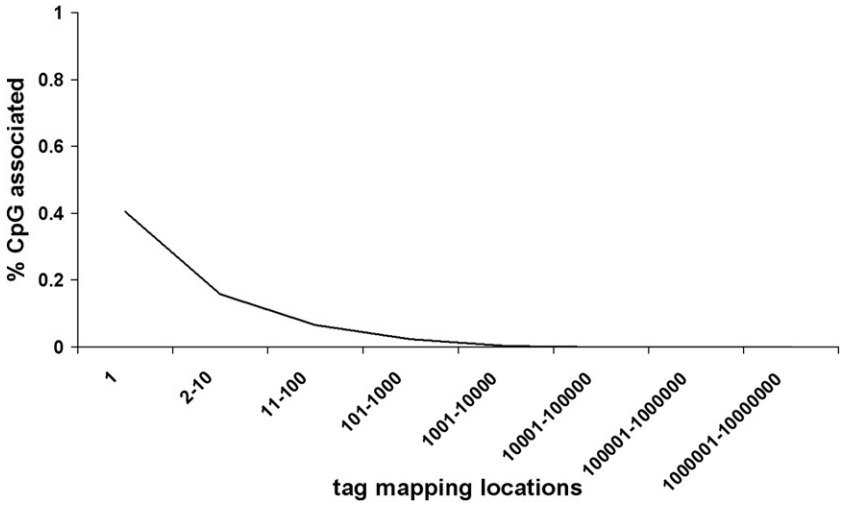

Fig. 2. Percentage of CpG-associated CAGE tags versus number of mapping locations.

locations that originated from within a repeat element was comparable to the percentage observed for uniquely mapped tags. From this we concluded that although repeat elements were clearly driving the production of "massively" multimap tags, the majority of multimap tags were not generated from within repeat elements.

Finally, a motif discovery algorithm based on position-weight matrices, MEME [19], was used to find statistically overrepresented motifs in the multimap tag set. To be considered as a major contributor to the generation of multimap tags, a motif had to occur in >10% of the sequences extracted for each category of log normalized mapping frequencies. Furthermore, only motifs longer than 8 nucleotides were considered. Using these criteria, the only motif found in any category was the dinucleotide repeat GAGAGAGAGA, which occurred in 14.1% of tags mapping to 101–1000 genomic locations. From this we concluded that the overwhelming majority of multimap tags were not associated with a widely represented, shared TSS-proximal motif.

Multimap tags are a consequence of redundancy in short genomic sequences

A previously unexamined explanation for the observed proportion of multimap CAGE tags was that they were simply a consequence of intrinsic redundancy in short sequences extracted from the mammalian transcriptome. To test this, the exonic, intronic, intergenic, and known promoter regions of the mouse genome, in addition to the 5′ ends of known terminal 5′ exonic sequences, were randomly sampled (n=100,000) and matched back to the genome to determine the expected multimap proportion for tag lengths of 12–30 nt.

Multimap rates for exonic, intronic, promoter, and initial 5′ transcript tag sequences were inversely related to tag length (Fig. 4), with approximately 20% of 20-mers mapping to more than one genomic location. However, when the mapping stringency for the 5′ transcript tag sequences was relaxed to allow a single mismatch the multimap rate increased to 45% for 20-mers.

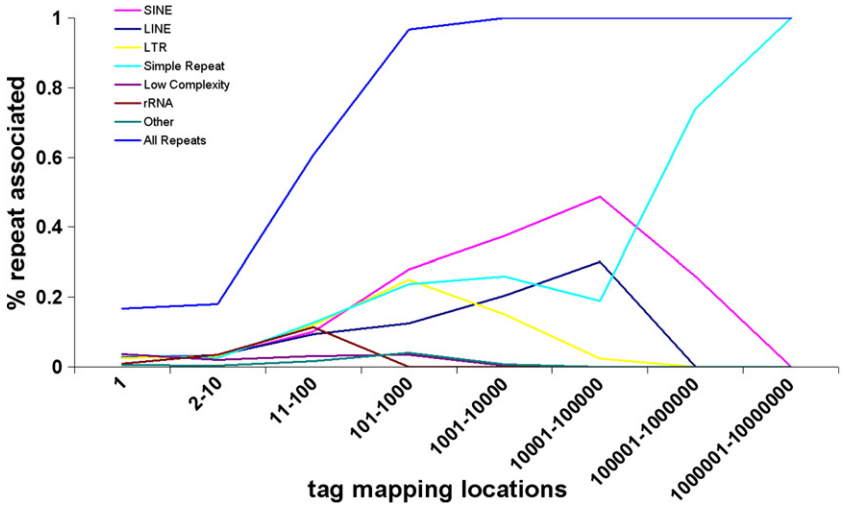

Fig. 3. Percentage of repeat element-associated CAGE tags versus number of mapping locations, by class.

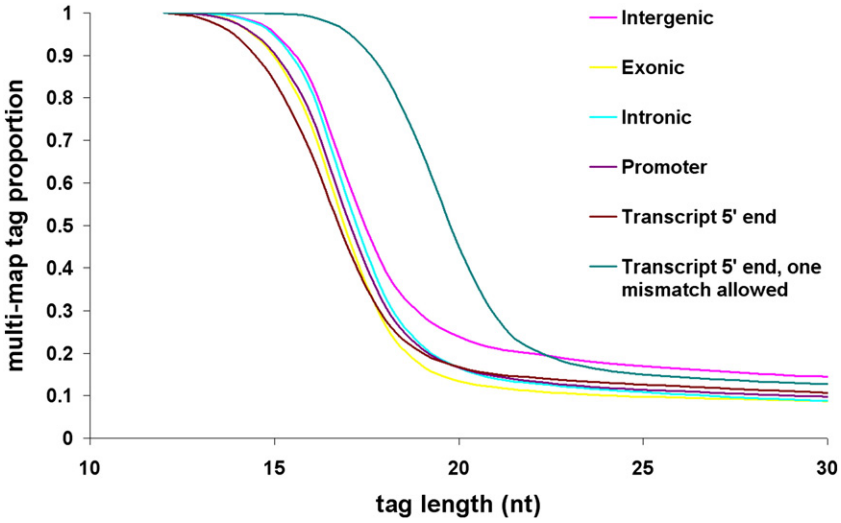

Fig. 4. Multimap proportions for random short-sequence tags in the mouse genome. Coordinates extracted from the UCSC genome browser [17] were used to identify each set of regions. Promoter sequences were taken at random from 1 kb upstream of transcripts. All sequences were matched perfectly to the genome, with the exception of the transcript 5′-end tags, which were mapped both perfectly and with one mismatch allowed. Each category consisted of 100,000 sequences for each tag length.

This predicted multimap rate was very similar to the observed fraction reported in Fig. 1 (41%). From this we concluded that redundancy in short transcriptomic sequences, compounded by a necessarily relaxed mapping strategy, was responsible for the vast majority of FANTOM3 multimap CAGE tags.

Paralogous promoters produce a minority of multimap tags

Gene duplication events were a highly plausible explanation for the observed level of redundancy in short transcriptomic sequences. By extension, the paralogous duplication of promoter regions could produce a large number of multimap CAGE tags. To address this theory we first calculated the rates at which tags that mapped to one to five locations were associated with known promoter regions by comparing their genomic coordinates with those of a clustered, nonredundant promoter set (see Materials and methods).

The proportion of mapping locations associated with a known promoter was highest for single-map tags (~70%) and decreased as a function of mapping location count to 13% for tags mapping to five locations (Fig. 5A). However, the rate at which tags were associated with at least one known promoter decreased far less quickly (47% for five locations). In other words, the bulk of multimap tags mapped to both a known promoter and multiple other genomic regions lacking prior evidence of transcriptional activity, suggesting that known paralogs contributed at most a substantial minority of multimap tags.

To confirm that the large number of previously unannotated mapping locations for multimap tags did not represent novel paralogs or pseudogenes [20] we calculated the average similarity between sequences flanking promoter-associated (PA) and non-promoter-associated (NPA) pairs. More specifically, this was achieved via pairwise BLAST [16] alignments between sequences extracted from −25,+50 in relation to the tag start sites for every possible pairwise combination of mapping locations (PA/PA, PA/NPA, NPA/NPA).

As can be seen in Fig. 5B, pairwise alignments for PA/PA tag location pairs could be extended on average between 30 and 40 bases beyond the boundaries of the tag coordinates, much farther than for PA/NPA and NPA/NPA pairs (5 and 10 bp, respectively). This implied that the regions immediately surrounding PA/NPA and NPA/NPA pairs lacked the similarity found for PA/PA pairs, suggesting that PA/NPA and NPA/NPA pairs did not in the main result from promoter duplications. Rather, the majority of multimap tags were associated with a single established promoter and multiple spurious genomic hits.

A novel rescue strategy for multimap tags

With these findings in mind, we next sought a method to resolve the most likely mapping locations for multimap FANTOM3 CAGE tags post hoc. To achieve this, we proportionately assigned tags that mapped to fewer than 100 genomic locations to the aforementioned clustered promoter sequences. Tags that mapped to more than 100 genomic locations were discarded because their contribution to the overall multimap tag complement was small and because any rescue of these tags would be tenuous at best.

The fundamental variability of transcription start site location, as captured by CAGE, proved crucial in this effort as it provided relationships between tags mapped to different positions in a common promoter. Every instance in which two independent tags mapped to the same promoter, and that was also the only instance of the two tags occurring in the same promoter, was counted as +1 to the score for each tag to be associated with that promoter. After every tag–promoter association had been processed, tags were proportionately assigned to their corresponding promoters based on the score for that particular association divided by the total score of all

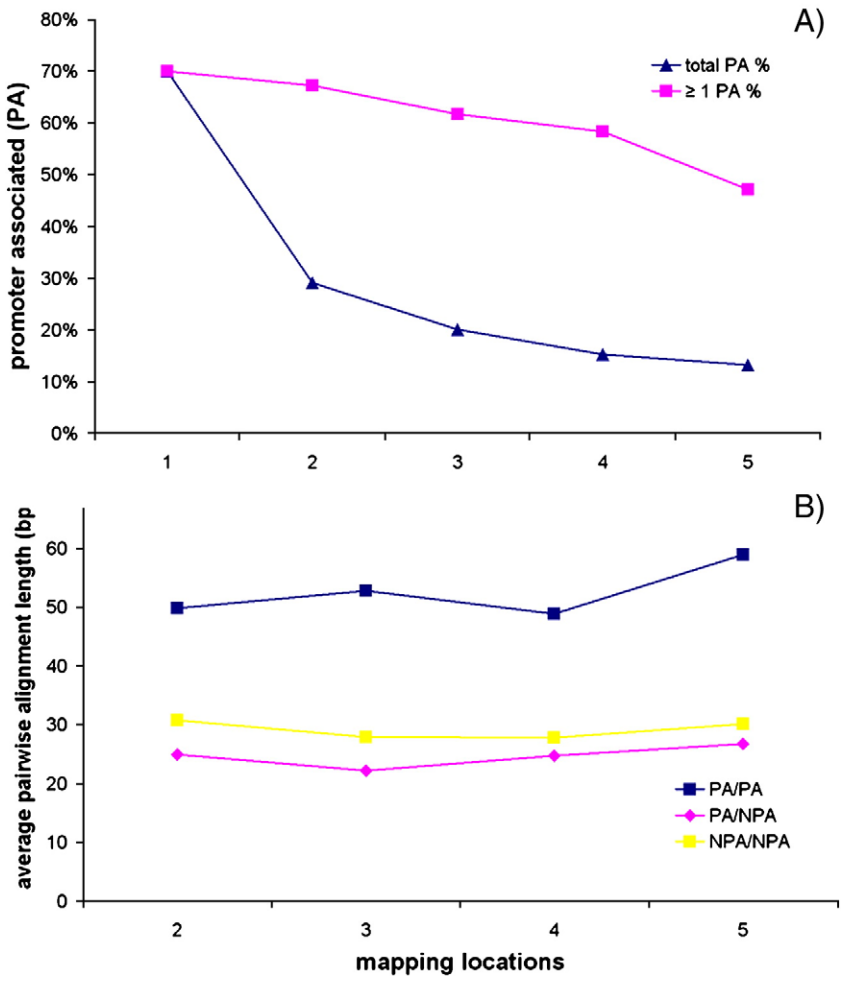

Fig. 5. Analysis of CAGE tag mapping to known promoters and other genomic locations. (A) Percentage of all multimap tag locations that correspond to a known promoter (total PA %, or total promoter-associated %) and percentage of tags that map to at least one known promoter (≥1 PA %), for tags mapping to one to five genomic locations. (B) Average length of pairwise BLAST alignments between promoter-associated/promoter-associated (PA), promoter-associated/non-promoter-associated (PA/NPA), and NPA/NPA pairs, for tags mapping to two to five genomic locations. For example, if a tag maps to three genomic locations, two of which are within known promoters, then it will produce one PA/PA pair and two PA/NPA pairs.

associations for that tag. This process is illustrated and explained further in Supplementary Fig. 1.

The selected rescue strategy had two main strengths: first, every tag–tag relationship for single-map tags was unique, meaning that single-map tags were implicitly assigned to a single promoter and that any multimap tag occurring proximal to a single-map tag was likely to be assigned to the same promoter as the single-map tag. Single-map tags therefore served to attract multimap tags to promoters.

Second, a rescue strategy based on splitting multimap tag counts over multiple promoters was analogous to a "rich get richer" approach; strongly expressed promoters were likely to supply a large number of unique tag–tag relationships and thereby attract more multimap tags. For instance, if a multimap tag occurred in two promoters at opposite ends of the expression scale, that multimap tag would most likely be associated with the highly expressed promoter. Promoters thought to be poorly expressed before application of the rescue strategy were therefore unlikely to switch to being highly expressed unless a large number of distinct multimap tags occurred together in only one common promoter.

Once this rescue strategy had been implemented we sought to evaluate its impact upon transcriptome coverage, expression profiling, and promoter classification by CAGE.

Improved transcriptome coverage via rescued multimap CAGE tags

As stated previously, a substantial proportion of the protein-coding loci defined by FANTOM3 could not be associated with a CAGE tag. To examine whether rescued multimap CAGE tags would resolve this shortfall, the variant transcript set (VTS)-derived promoter sequences defined under Materials and methods were categorized based on whether each was detected by more than one (a) single-map tag or (b) single-map *or* rescued multimap tag.

A minority of the promoter sequences were detected by more than one single-map tag, with coverage for protein-coding and noncoding promoters of 43.8 and 21.4%, respectively (Fig. 6). The corresponding percentages for single *or* rescued multimap tags were 55.7 (a difference of 5975 promoters) and 40.1%

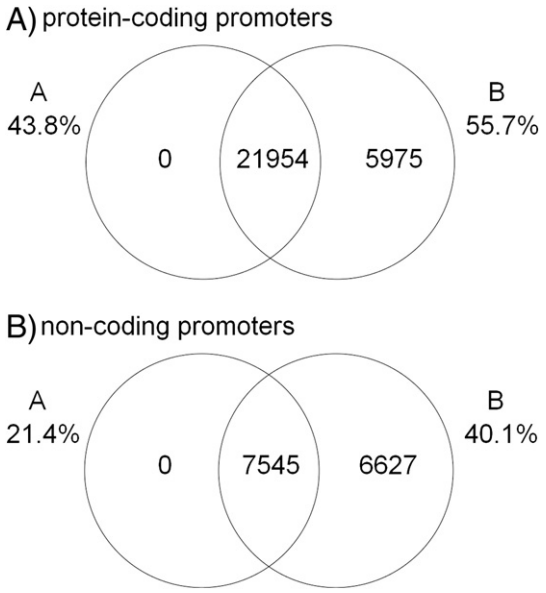

Fig. 6. Transcriptome coverage by CAGE. Promoter regions derived from VTS transcripts were associated with (A) >1 uniquely mapped CAGE tag or (B) >1 rescued CAGE tag for (a) protein-coding promoters and (b) noncoding promoters. The given percentages indicate the proportion of protein-coding and noncoding mouse promoters covered by A or B.

(6627 additional promoters). In reference to TKs, protein-coding coverage increased from 42.8 to 57.8% (5516 more frameworks) and noncoding coverage improved from 8.0 to 32.3% (6082 additional frameworks). If the threshold for coverage was lowered from "more than one tag count" to "one or more tag counts"—to include singleton tags—coverage improved to 61.2 and 50.1% for coding and noncoding promoters, respectively.

As indicated in Fig. 6, the majority of detected promoters contained at least one redundant subregion. Indeed, by generating every possible 21-mer within a set distance of each TSS, 13.6% of all transcripts were expected to produce only multimap tags based on the 25-nt region either side of their annotated TSS, 9.2% if 50 nt on either side was considered, 6.0% if 100 nt on either side was considered, and 4.5% if the entire −300,+100 promoter regions were considered. Between 4.5 and 13.6% of transcripts were therefore considered unable to generate single-map tags.

When viewed from an overall perspective these coverage statistics indicate three areas in which the transcriptome-wide application of CAGE can be improved. First, the substantial gap in transcriptome coverage by CAGE, even after the inclusion of rescued multimap tags, suggests that complete transcriptome resolution will require more tissues and activation states than were surveyed by FANTOM3, particularly if the increasing number of known tissue-specific and inducible promoters is considered [21,22]. Second, the additional coverage permitted by singletons suggests that many transcripts will be reliably detected only through deeper CAGE sequencing than what was feasible for FANTOM3. Last, multimap tag rescue strategies, such as the one detailed here, will be required to examine highly redundant promoters even with the introduction of more libraries or deeper sequencing.

Rescued multimap CAGE tags improve cross-platform concordance with microarray data

Given the prevalence of mixed single-map/multimap tag promoters (as outlined in Fig. 6), we hypothesized that past expression profiling based exclusively upon single-map tags was skewed by the omission of multimap tags. To test this, a cross-platform comparison of FANTOM3 CAGE macrophage data with Affymetrix GNF array data [23] that incorporated both single-map and "rescued" multimap tags was performed.

More specifically, expression ratios comparing unstimulated macrophages with lipopolysaccharide (LPS)-treated macrophages were calculated for every equivalent promoter–GNF probe pair. These ratios were then inserted into the conservative Up/Down method (see Materials and methods for further details) to quantify cross-platform concordance [24]. Given the technological differences between CAGE and microarrays (5′ transcript end versus 3′ end, sequence tag based versus hybridization based) a moderate correlation was expected.

This prediction was confirmed by a correlation of 0.62 between platforms based on single-map tags only. Although this improved slightly to 0.67 once rescued multimap tags were included, the biggest difference between the two approaches was that the use of multimap tags ($n=123$) permitted nearly 50% more comparisons than the use of single-map tags alone ($n=84$). From this it was concluded that the multimap tag rescue strategy mildly improved the accuracy and dramatically expanded the coverage of the FANTOM3 CAGE set when applied to expression profiling.

Promoter reclassification using rescued multimap CAGE tags

The CAGE-based promoter classification approach detailed by Carninci et al. [11] was revised to include the rescued multimap CAGE tag set. To summarize, in the original publication mammalian promoters were grouped into four main classes based on the distribution of CAGE tags around their annotated TSS. Briefly, these classes were single dominant peak (SP), general broad distribution (BR), broad distribution with a dominant peak (PB), and bi-or multimodal distribution (MU). Using only single-map tags for promoters with abundant CAGE support (>100 tags), Carninci et al. concluded that the SP, TATA-box-

Table 1
Promoter classification differences due to the introduction of multimap tags

Class	Original ($n=5872$)	Rescued ($n=2567$)	Total ($n=8439$)
SP	11.8 (50.3)	17.0 (57.4)	13.8 (53.0)
PB	51.4 (11.7)	43.0 (21.8)	49.2 (14.3)
MU	3.7 (12.9)	2.9 (17.1)	3.5 (14.0)
BR	33.0 (5.5)	37.0 (9.3)	34.3 (6.8)

Proportions for the SP (single dominant peak), BR (general broad distribution), PB (broad distribution with a dominant peak), and (MU) bi-or multimodal distribution promoters are detailed for promoters with >100 single-map CAGE tags (Original), promoters with >100 tags after multimap rescue and <100 tags prior to rescue (Rescued), and for all cases (Total). Also given in parentheses after each value is the proportion of promoters with a TATA box within 50 bp of their strongest peak. The promoter set described here was previously used to generate Fig. 6 (bases −300 to +100 extracted, then clustered).

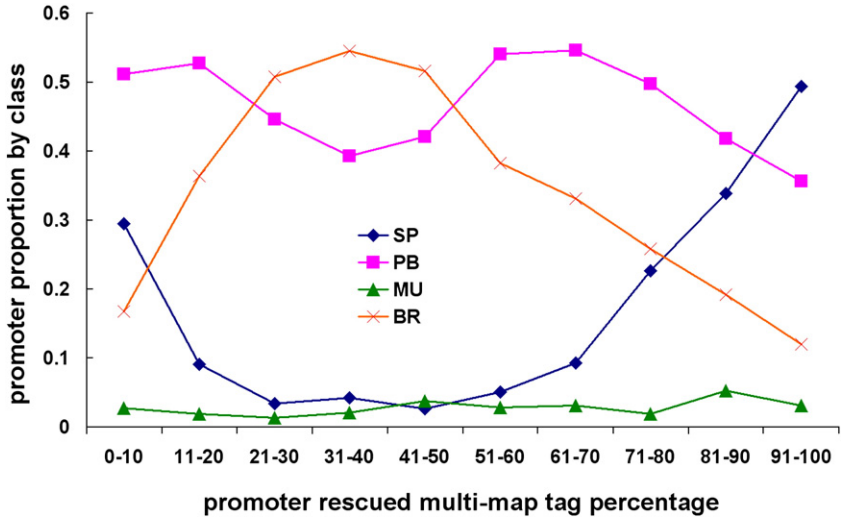

Fig. 7. Promoter class proportions compared with promoter multimap tag percentage after rescue. All promoters had more than 100 tags after the multimap tag rescue.

associated, promoter architecture occurred relatively infrequently in mammals.

When we performed a similar analysis using rescued multimap tags the number of promoters matched by >100 tags increased from 5972 to 8439, with minor increases in the overall incidence of SP promoters and their TATA-box association rate (Table 1). However, it was noticed that promoters supported by >90% rescued multimap tags possessed an SP architecture in approximately 50% of cases (Fig. 7), compared to 13.8% overall.

This observation can be explained probabilistically, for if 40% of CAGE tags multimapped then we would anticipate that the tag count per promoter would be reduced by 40% on average upon the omission of multimap CAGE tags. However, this reduction would affect SP and non-SP promoters very differently; the broad and robust nature of non-SP promoters would lead to very few of this promoter class becoming undetected when multimap tags are removed. On the other hand, ~40% of SP promoters would have virtually all of their tags removed and ~60% would be unaffected because these promoters rely on one or two positions to supply the vast majority of their tags. A promoter comprising >90% rescued multimap tags is therefore far more likely to present an SP architecture than a promoter supported by a low proportion of rescued multimap tags. These SP promoters in particular would not have been detected without the multimap CAGE tag rescue strategy presented here.

Conclusions

A systematic review of the FANTOM3 mouse CAGE tag set revealed that the most likely cause of multimap tags was a high level of redundancy in short sequences extracted from transcribed regions of the genome, compounded by a necessarily relaxed mapping strategy. A substantial portion of the multimap tag complement was subsequently rescued, leading to expanded CAGE transcriptome coverage, improved cross-platform agreement with array data, and the discovery that a disproportionate number of promoters largely supported by rescued multimap tags conformed to a narrow architecture. We believe these applications, combined with the rapid development of other high-throughput sequencing methodologies based on short sequence reads, emphasize the importance of a multimap tag rescue strategy.

Materials and methods

CAGE tag mapping

Each mouse FANTOM3 CAGE tag was matched against MM8 using Vmatch, with a minimum match length of 18 bp, a single internal mismatch allowed, and multiple mismatches allowed at tag ends. For each tag, the alignment providing the highest number of correctly aligned bases was selected as the best match. If a tag matched equally well to more than one genomic location it was designated a multimap tag. The same criteria were used to map tags to promoter sequences.

Generation of clustered, nonredundant promoter sequences

Promoter sequences (base positions −300 to +100) were extracted for every unique TSS defined by the mouse VTS (a comprehensive, nonredundant set of full-length transcripts [25]). Promoter regions that overlapped by more than 200 nt were joined into a single cluster.

Cross-platform expression profile comparison

To begin with, Affymetrix GNF probe sequences were matched to VTS transcripts using Vmatch, with no mismatches allowed and every subprobe of each GNF probe set required to match to a transcript for a GNF probe set–transcript match to be recorded. If the list of transcripts matched by a given probe set was identical to the list of transcripts represented by a clustered promoter sequence the two elements were considered comparable (i.e., CAGE and GNF were measuring the same set of transcripts).

After the rescue strategy had been implemented, single-map and rescued multimap CAGE tags from libraries surveying unstimulated and LPS-treated (7-h time point) macrophages were used to assign promoter expression levels. GC Robust Multiarray Average (gcRMA) [26] normalized GNF expression data for the equivalent macrophage time points (presented at the Symatlas Web site, http://wombat.gnf.org/SymAtlas/; GNF Expression Data Viewer) were then obtained. If two GNF probes occurred in a common cluster of transcripts their signals were averaged. At this point we had comparable promoters and GNF expression data, with promoter expression levels based on both single and rescued multimap tags.

Next, to circumvent the problematic normalization necessary to compare tag-based expression data directly with array data, we calculated a ratio (unstimulated macrophage library CAGE tag count/7-h LPS-treated macrophage library CAGE tag count) for promoters and a ratio (unstimulated macrophage signal intensity/7-h LPS-treated macrophage signal intensity) for GNF probes. These ratios were then inserted into the Up/Down correlation method, as discussed by Van Ruissen et al. [24].

Eight hundred twenty-seven promoters were detected by at least one single or rescued multimap CAGE tag and had an equivalent GNF probe. Four hundred three GNF probes produced at least one gcRMA normalized signal intensity >200 relative fluorescence units, while 508 promoters could be associated with more than one CAGE tag. To further filtering, only those promoters and probes that changed more than 10% (i.e., a ratio less than 0.9 or greater than 1.1) were compared, leaving 84 promoter–probe pairs based on single-map tags and 123 promoter–probe pairs after the addition of multimap tags.

Up/Down correlations for comparisons based on single-map or single-and rescued multimap tags were then calculated by dividing the number of comparisons in which the promoter and probe ratios were both above 1.1, or both below 0.9, by the total number of comparisons (as illustrated in Supplementary Fig. 2). Correlations of 0.62 and 0.67 were noted for comparisons based on single-map or single-and rescued multimap tags, respectively.

Specifications for promoter classification

To determine the relative frequencies of each promoter class we defined a small peak as a single promoter position that contained >30% of the total tags for a promoter, and a large peak contained >80% of the tags for a promoter. The SP, PB, MU, and BR classes were then defined as follows: (1) SP, one large peak; (2) PB, no large peak, one small peak; (3) MU, no large peak, more than one small peak; (4) BR, no large or small peaks.

Acknowledgments

G.F. was supported by an Australian Postgraduate Award through the Australian Government Department of Education, Training and Youth Affairs. A.R.R.F. is funded by a C.J. Martin Fellowship from the Australian NHMRC (ID 428261). K.S. is a member of the CRC for Chronic Inflammatory Diseases. A.M.C. was supported by the Queensland Government Smart State Initiative and the Australian Research Council (DP0665078). P.C. and Y.H. were supported by the National Project on Protein Structural and Functional Analysis from MEXT and the National Project on Genome Network Analysis and the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science, and Technology of the Japanese Government. D.A.H. acknowledges the funding of the ARC Special Research Centre for Functional and Applied Genomics and the NHMRC. S.M.G. was supported by the ARC Centre in Bioinformatics, the ARC Special Research Centre for Functional and Applied Genomics, and an NHMRC Career Development Award.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.11.003.

References

[1] P. Carninci, et al., The transcriptional landscape of the mammalian genome, Science 309 (2005) 1559–1563.
[2] P. Carninci, et al., Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia, Genome Res. 13 (2003) 1273–1289.
[3] Y. Hayashizaki, P. Carninci, Genome Network and FANTOM3: assessing the complexity of the transcriptome, PLoS Genet. 2 (2006) e63.
[4] J. Kawai, et al., Functional annotation of a full-length mouse cDNA collection, Nature 409 (2001) 685–690.
[5] Y. Okazaki, et al., Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, Nature 420 (2002) 563–573.
[6] P. Carninci, Y. Hayashizaki, High-efficiency full-length cDNA cloning, Methods Enzymol. 303 (1999) 19–44.
[7] P. Carninci, et al., High-efficiency full-length cDNA cloning by biotinylated CAP trapper, Genomics 37 (1996) 327–336.
[8] P. Carninci, et al., Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes, Genome Res. 10 (2000) 1617–1630.
[9] T. Shiraki, et al., Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 15776–15781.
[10] V.B. Bajic, et al., Mice and men: their promoter properties, PLoS Genet. 2 (2006) e54.
[11] P. Carninci, et al., Genome-wide analysis of mammalian promoter architecture and evolution, Nat. Genet. 38 (2006) 626–635.
[12] P.G. Engstrom, et al., Complex loci in human and mouse genomes, PLoS Genet. 2 (2006) e47.
[13] M.C. Frith, et al., Evolutionary turnover of mammalian transcription start sites, Genome Res. 16 (2006) 713–722.
[14] A. Sandelin, et al., Mammalian RNA polymerase II core promoters: insights from genome-wide studies, Nat. Rev., Genet. 8 (2007) 424–436.
[15] S. Kurtz, The Vmatch large scale sequence analysis software, 2006.
[16] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[17] W.J. Kent, et al., The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.
[18] A.F.A. Smit, G. P., RepeatMasker.
[19] T.L. Bailey, N. Williams, C. Misleh, W.W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Res. 34 (2006) W369–W373.
[20] E. Birney, et al., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447 (2007) 799–816.
[21] A.R. Forrest, et al., Genome-wide review of transcriptional complexity in mouse protein kinases and phosphatases, Genome Biol. 7 (2006) R5.
[22] C.A. Wells, et al., Alternate transcription of the Toll-like receptor signaling cascade, Genome Biol. 7 (2006) R10.
[23] A.I. Su, et al., A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 6062–6067.
[24] F. van Ruissen, et al., Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips, BMC Genomics 6 (2005) 91.
[25] T. Kasukawa, et al., Construction of representative transcript and protein sets of human, mouse, and rat as a platform for their transcriptome and proteome analysis, Genomics 84 (2004) 913–921.
[26] R.A. Irizarry, et al., Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Res. 31 (2003) e15.